DOCUMENT RESUME

ED 229 391 TM 830 226

AUTHOR Gullickson, Arlen R.
TITLE The Practice of Testing in Elementary and Secondary Schools.
PUB DATE 16 Nov 82
NOTE 18p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Elementary School Teachers; *Elementary Secondary Education; Secondary School Teachers; Teacher Education; *Teacher Made Tests; Test Construction; Test Format; *Testing; *Testing Problems; Test Results; Test Theory; *Test Use

ABSTRACT

Rudman and colleagues (1980) deplored the paucity of descriptive information relative to teachers' test use patterns. The present study addresses the abundant prescriptive, and lack of descriptive information concerning teacher testing. A mailed survey procedure gathered testing practice information from elementary and secondary South Dakota school teachers (n=336) regarding: (1) testing context, (2) test construction, (3) test administration, (4) test analyses, and (5) test results. The survey indicated that teachers use a variety of testing techniques, but only teacher-made objective tests play a major evaluative role across all grade levels and curricular areas. There appear to be three important factors which influence teacher practice: time, expertise, and tools available for teachers' use. Nearly 20 percent of in-class time is devoted to test-related activities. This substantial time investment is a strong argument for skill in the practice of testing; however, most teachers have limited preparation in the area. Improved practices require changing the habits of teachers and educating them to overcome their lack of knowledge of sophisticated tools (e.g., calculators, microcomputers). Perhaps the most clear need is for a return to development of measurement techniques that will be appropriately used in the classroom. (Author/PN)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED229391

The Practice of Testing in Elementary and Secondary Schools

Gullickson, Arlen R.

Educational measurement issues which reach us in overt ways, e.g., through the press, typically deal with standardized measures of aptitude and achievement. Yet the most pervasive use of measurement occurs in the context of normal classroom routine. Such measurement through formal and informal assessment processes, forms an important basis of communication among teacher, student, and parents. This communication tends to be personal, not public, low profile, i.e., not involving or engendering public discussion, and is controlled by the teacher. Because the communication has those characteristics its measurement basis is rarely subject to close scrutiny.

What are the measurement practices of teachers? More specifically, what is the context in which tests are given; how are tests constructed, administered, analyzed, and reported? These are questions pertinent to the improvement of testing practices; questions which teachers might ask themselves in self reflection; and questions which measurement specialists must address in helping teachers to use tests effectively.

Measurement specialists (cf. Hopkins and Stanley, 1981) view evaluation processes as interacting with educational objectives and learning experiences which together comprise the educational process. Whether evaluation processes, in

TM 830 226

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
A. R. Gullickson

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

particular tests, actually do function in this manner is open to question. Rudman and his colleagues (IRT, 1980, p.20), after a review of literature covering nearly 60 years, deplored the paucity of descriptive information relative to teachers' test use patterns. Their review makes it clear that while prescriptive information is abundant, the lack of descriptive data makes it impossible to determine if the prescriptions fit, are appropriate to, practice. This study was initiated to address that prescription/practice gap and focused on the teacher testing practice questions posed earlier.

A mailed survey procedure was used to gather the information from teachers who were sampled from the South Dakota directory of teachers in elementary and secondary schools. In all 75%, 336 of a total of 450 teachers, stratified by grade level (grades 3,7,and 10) and curricular area (science, social science, and language arts), responded to the questionnaire. In each case the cover letter asked the teacher to respond relative to personal testing practices.

The teachers who responded to the questionnaire appear to be typical of teachers in general. That is, they are college graduates holding at least a bachelor's degree with a quarter (24%) having a master's or higher level degree. They are experienced teachers the majority (50%) having taught 10 years or more. Ninety-five percent teach at least three classes a day and the majority have at least three

course preparations. The majority have taken only one course (57%) or no course (5%) in educational measurements, but for a large majority (84%) other courses have provided some information about the preparation of tests.

Almost all of these teachers use tests with 89% of the elementary teacher and 99% of the secondary teachers (junior and senior high school) reporting such use. Not only do they test, but they do so frequently. Virtually all test on a weekly (95%) or at least a biweekly (98%) basis. In this testing process they use a variety of testing techniques, but only teacher-made objective tests play a major evaluative role across all grade levels and curricular areas.

The questionnaire to which teachers responded was built on the premise that test use is cyclical in nature. That is, a test is initiated to meet specified purposes; preparations are made, the test is administered and analyzed, and the results are used in the context of intended purposes. Thus in responding teachers first provided contextual information. Then in the order cited above they answered items regarding their personal testing practices.

RESULTS

Responses were analyzed by grade and curricular level to identify practices which are related to those two variables. Where significant effects were found, they are reported. In those situations where the dependent variables had interval scale characteristics, and several dependent variables were

analyzed together, multiple analysis of variance techniques were employed (SAS, 1979). Where individual dependent variables were analyzed, if the dependent variable had interval scale characteristics, analysis of variance techniques were used. If only frequency counts were available for the dependent variables, Chi-square and contingency table analyses were conducted.

The Testing Context

When queried as to the role that several different types of tests had in their evaluation of students, teachers reported teacher-made objective tests as having the greatest role, essay tests as having the second largest role, followed by standardized objective tests and oral quizzes. Of the four, objective tests received much higher ratings than did all of the other three. Essay tests received high ratings at the secondary level but very low ratings at the elementary level. In general, the role of testing in the classroom increases from the elementary to the secondary level (Note 1). The role of testing also differs significantly but not substantially across curricula.

Testing is a time consuming activity. For example, in the use of teacher-made tests, some teachers report spending more than nine hours per individual test in the various test related activities. The typical, i.e. median teacher, reported spending slightly over three hours (190 minutes) on

test related activities. Roughly this breaks out to 60 minutes for test preparation, 30 minutes for test correction and 20 minutes for post test review.

Given this background of teacher experience, the role of testing for teachers, and the amount of time teachers spend in the context of testing, teachers were asked which of several purposes classroom tests were expected to fill. Six separate purposes were identified and for each of those purposes the teacher was asked to rate on a four-point basis which constituted the purposes for which they used classroom tests (0 = not a purpose, 1 = minor, up to 3 = major purpose of the test). Of the six, three received mean ratings of approximately 2.6. These were: instructional feedback for student learning (2.64), evaluation of instruction (2.62), and evaluation (grading) of students (2.58). Motivation of student study ranked fourth in ratings (2.23). The remaining two, assessment of the attitudes or interests of students (1.54) and providing opportunity for student input into evaluation of instruction (1.47) received substantially lower ratings.

Two, evaluation or grading of students and the assessment of attitudes and interest varied by grade level. Teachers placed less emphasis on grading purposes at the elementary level and progressively more emphasis through the senior high. The mean rating at the elementary level was 2.34 with a mean rating of 2.7 at the senior high level.

Mean ratings on the assessment on the attitudes and interests of students moved in just the opposite direction, being highest at the elementary level (1.81) and much lower at the secondary level, i.e., 1.36 for junior high and 1.46 for senior high, respectively. Clearly, teachers perceive tests as serving an instructional purpose both for feedback to the students and feedback to the teacher with grading of students maintaining an important role in that feedback.

Test Construction

Teachers were asked about their source of items, the types of items that they constructed, whether or not their tests covered all of the material they teach, and whether or not they reuse their tests in subsequent semesters. They identified two primary sources for items. First, 93% view themselves as a primary source of items, i.e. they write their own items. Second, 60% also report using test items prepared by the publisher of the textbook which they are using. Two other sources, other published test items and test items prepared by other teachers were identified as primary sources by substantially fewer teachers (23% and 12%, respectively).

In three of these areas, there were grade level or curricular area differences. A slightly lower percentage of elementary teachers write their own items (85%) as opposed to 96% for the secondary teachers. Also, elementary teachers

are more prone to use textbook items, 75% vs. 61% and 47% for elementary, junior high and senior high school, respectively. Third, although items prepared by other teachers were not the primary source for most teachers, junior high teachers were more prone to share items than were other teachers, i.e. 20% for junior high vs. 9% for elementary or senior high teachers.

When asked to identify what types of items they normally construct, most (54%) teachers checked several, i.e. four to five separate item types. The most popular type of item was short answer/completion (92%) followed by matching and multiple-choice (77% and 76% respectively), followed by true/false (68%), and finally, essay items (58%). The use of multiple-choice and essay items both differed significantly across grade level with fewer teachers at the elementary levels choosing those items than at the secondary level. The use of true/false items differed across curricular area with more teachers in the social sciences choosing to use true/false items than teachers in either science or language arts areas (83% vs. 69% and 55% for social science, science and language arts, respectively).

Two other general perspectives of test preparation were provided. One, even though teachers prepare their own tests they do not perceive the test as adequately evaluating all that they teach. Rather, the average teacher perceives tests to cover approximately 75% of the material taught. Second,

once teachers have prepared tests they tend to reuse those tests in the future. Eight-four percent of the teachers report reuse of their tests of which 60% report reusing all or major parts of the tests and 25% report reusing selected items. Thus, for most, the preparation of the test does not require totally constructing a new test each time a test is administered.

Test Administration

Testing appears to be a formal, constrained situation in which students expect to be graded. Virtually all teachers (99%) do not allow student interaction during the testing process. A substantial percentage, 26%, do not even allow students to ask questions of the teacher. In addition students are constrained in their use of support material. Seventy-nine percent of the teachers do not allow students to use their textbook, notes, etc., in completing a test. An exception to this general statement on support materials occurs in the use of calculators. While in general, 89% of the teachers do not use the calculators, in the area of science where calculator use might be most prominent, 40% of senior high school teachers allow use during tests. (It seems likely that this percentage would be substantially higher if teachers of physics and chemistry in grades 11 and 12 were queried.)

Teachers were also asked whether or not students were to provide answers in the test booklet or on separate answer sheets. While most require students to answer in the test booklet, a substantial minority, 36%, do require the use of a separate answer sheet. This seemed important from two perspectives. First, if the tests were speeded, that is given within time constraints such that many students could not finish the test, the use of separate answer sheets would be a substantial concern. Second, the use of a separate answer sheet provides opportunity for test booklets to be reused.

Speededness of tests appears not to be a problem as most, 92%, provide sufficient time such that almost all students have as much time as they need to finish the test. Regarding potential reuse, of those which use separate answer sheets 17% say it is solely for reuse of the test booklet, 38% say such use is solely for administrative ease in scoring, and 24% say it is for both administrative ease in scoring and reuse of the test booklet. Thus, approximately 20% of all teachers set up the administration of the test so that scoring of the test is facilitated and approximately 10% of the teachers set up test administration procedures to facilitate future reuse of the test booklets.

Test Analyses

Teachers were asked to rate on a four-point basis --always, usually, sometimes, and never--, to respond to several items with regard to their scoring and grading practices. Here teachers report that they do their own scoring and grading of tests, i.e. 95 to 97%, respectively, either always or usually grade their own tests. (Junior high teachers report being slightly less likely to score and grade their own tests than are elementary and senior high teachers.)

Typically, teachers assign a test grade rather than providing only a numerical score. In this context, rarely do teachers just assign a pass/fail grade to student tests, (mean score of 3.7 where 3.0 equals sometimes and 4.0 equals never). Related to this, most teachers (78%) use a criterion reference scheme for grading tests; only 10% use a curve basis for grading. Here criterion reference was used in the context of the example 90% or better for an A, 80 - 90% B, etc. In addition to scoring and grading tests, 90% report providing written comments to students on at least an occasional basis, with 55% reporting they always or usually do.

A second set of items asked teachers to identify which statistics they used in working with test results. Here teachers report using relatively little statistical information. Ninety percent indicate that they provide a total test score. Less than half, 42%, obtain the range of test scores. Few, 10% to 13%, use such information as the mean, median

and standard deviation. A fairly large minority of teachers reported use of item difficulty and reliability information, 31% and 29% respectively.

Clearly many teachers erred in checking reliability and item difficulty. For example, not only would it be unwise to talk about the reliability of the test without gaining information about the variability (standard deviation) of test scores, but calculation of the reliability requires knowledge of the standard deviation. Also, calculation of item difficulty, i.e. the percent of correct responses for each item, requires substantially more effort than does calculation of the mean, median or standard deviation. Thus, the high percent of response for item difficulty and reliability suggest that many teachers do not have an adequate understanding of either the terms or how such information is obtained from test results.

Use of Test Results

Teachers attempt to return test results to the students in a timely manner; only 6% required more than two days to process tests for return, 83% returned the tests within one day of the test, and 7% indicate that they return the tests the same day.

Teachers were asked to apportion time spent with the class in review of the test into three categories: 1) review of scoring and grading procedures, 2) review of individual

test items based on individual students request, and 3) review of individual items based upon the teacher's perusal of student results. The average techer indicates that 16% of the time is spent in review of grading procedures, 41% in the review of individual test items based on student requests, and 43% in the review of items and item groups based on teacher perusal of tests. When viewed in the context of a median total time of 20 minutes spent in the class review of tests, this breaks down to approximately 9 minutes spent on items chosen by the teacher, 8 minutes on items chosen by the students, and three minutes spent in the review of grading procedures.

Finally, teachers were asked whether students were allowed to keep their tests, if tests were returned to the teacher and thus not available for individual student review, or whether tests were retained by the teacher and students were allowed to review the tests under supervision. In each case the teacher was to respond on an always, usually, sometimes, or never basis. Here, as might be expected, there were significant differences across grade levels. At the elementary level, the average teacher "usually" let students keep the tests. At the secondary level, teachers only "sometimes" let the students keep the tests. Commensurate with those findings, secondary level teachers more frequently retain the tests but allow students to review the tests under supervision.

A significant proportion of class time and teacher time is devoted to the activity of testing. If one estimates an overall average of 45 minutes per day, five days a week, is given to each class, and if it is also estimated that a teacher-made objective test is administered every other week, then nearly 20% of in-class time is devoted to test related activities. Probably even a higher percentage of total teacher work time is given to test activities. This substantial time investment is a strong argument for requiring teachers skilled in the practice of testing, and for developing efficient testing techniques.

But, as the results show, most teachers have very limited preparation in the area of testing. In the state of South Dakota, for example, collegiate programs routinely provide two semester credit hours of educational measurements to meet certification requirements. Any additional test information is provided at the discretion of individual faculty in methods courses. Other results suggest this limited educational experience is inadequate.

There are at least three tentative indicators that whatever is taught in pre-service courses does not spill over into sound testing practices by teachers. First, in the preparation of tests, short answer and matching items are the most popular items of choice. Both types tend to be limited to lower cognitive level, i.e. knowledge level, assessment

(Hopkins and Stanley, 1981). Thus tests probably assess only lower cognitive level understandings.

Second, while the large majority of teachers reuse items, few teachers take the time or make the effort to systematically improve their items. This is suggested by the minimal amount of time given to test analysis (barely enough to score and grade tests) and by the minimal use of test statistics. As a direct result, test item improvement must be done on a very ad hoc and subjective basis.

Third, teachers appear to misuse criterion referenced tests. On the surface teachers' advocacy of criterion referenced testing would indicate evidence of a firm criterion referenced testing foundation. However, even if teachers clearly define their test domain -- a topic not addressed in this survey -- they clearly do not address quality of items in a manner which would insure their items function as desired. Most reuse their items but without careful item analysis. Thus, criteria established by teachers are both artificial and subjective. For without knowing how items function, it is not possible to accurately set criterion levels for student performance. Regardless of the domain being tested, a test may be prepared which is very difficult or very easy. Also, the cognitive level of the test may be shifted so that only knowledge level items are addressed or higher cognitive level items are addressed as well. Results of this study suggest that neither test difficulty nor the

cognitive level of items has been adequately addressed by teachers, thus criterion referenced testing is simply a word and not an accomplished fact.

Potentially, the consequences of these concerns are substantial. If tests are oriented toward lower cognitive levels and students are graded on their attainment of such knowledge, students must be motivated to focus on lower cognitive level learning. Also, because teachers grade on a criterion referenced basis but without a-priori knowledge of how their tests will function, their expectations of students and their rewarding of students, grades, praise, etc., must vary as a function of test quality. Such testing effects seem undesirable!

There appear to be three important factors which influence teacher practice. They are time, expertise, and tools available for the teacher's use. Given the already substantial amount of time that teachers apply to the testing practice, it seems unlikely that teachers can substantially add to the amount of time presently being used. Thus, teachers must either reorient their time (for example by using less time in test preparation and more time in test analysis), or they must find more efficient methods for handling the process of testing.

Quite likely if the testing routines of teachers were studied in depth, there would be numerous ways of simplifying the testing practice process and improving its efficien-

cy. These new techniques could then be brought to teachers through in-service and pre-service instruction to improve teacher knowledge and effectiveness in the area of testing. Such efforts alone would not be sufficient. There remain a substantial proportion of the teachers who are either uninformed or misinformed about basic testing concepts, e.g. reliability. Such concepts need to be re-presented to teachers in ways which are compatible with their testing situation so that conceptual concept understanding grows rather than deteriorates over time.

The use of tools available to the teacher is the third area that seems very appropriate to pursue. While at first glance it would appear that the tools available to those in testing have remained constant over the past years, in fact a number of substantial changes have been made. For example, the advent of the photo copy machine essentially eliminates the need to retype an item each time it is used. The hand-held calculator makes computation of means, standard deviations, and even reliabilities a relatively straight forward and short process. Also, the microcomputer is sure to facilitate the development of items and item analyses, as well as the individual testing of students.

Personal experience suggests that it is a rare teach who stores items in a manner which allows test preparation without the need for individual item typing. Also, most teachers are relatively unfamiliar with the more sophisti-

cated calculators which can do means, standard deviations, and reliabilities in a straight forward manner and they are unfamiliar with the possibilities which exist in microcomputers. Thus, improved practices requires changing the habits of teachers, and educating them to overcome their lack of knowledge and fear of the more sophisticated tools. Even then teachers may need to be pursuaded that the payoff from improved tests is commensurate with the added effort.

If teachers are to improve their testing habits, and it seems important that they do, they will need assistance. This entails practical help in making them more efficient in their daily testing habits and new ideas and expertise in testing. Perhaps what is most clear is the need to return to the basics of measurement. That is, a return to development of measurement techniques that will be appropriately used in the classroom.

Reference List

Hopkins, K. D. & Stanley, J. C. Educational and Psychological Measurement and Evaluation (6th ed.). Englewood Cliff, NJ: Prentice Hall, 1981.

Institute of Research on Teaching. Integrating Assessment with Instruction a Review (1922-1980). College of Education, Michigan State University, 1980.